机器人监督员: 非侵入式算法公平性证明机制

林彦熹 *‡ 高航 *¶ 刘冬梅 § 于洋 †‡ 杜睿桓 †¶ 陆敏 ¶ 徐哲 ¶
郭杰成 ¶ 乔国梁 § 吴国斌 ¶

摘要

网约车平台中，订单分配算法作为其提供服务的基础设施撮合司机与乘客，直接影响司机与乘客的权益与出行体验，如何检验和确保订单分配算法的公平性对网约车平台良序运营、保障司乘两方权益是一个至关重要的话题。本研究提出一个非侵入式的网约车订单分配算法的公平性检验方法，无需深入算法细节，只需利用订单分配的场景信息和分配结果便能进行算法公平性检验，大幅降低计算成本，并能保护企业机密及用户隐私。本研究提出的方法首先通过统计建模随机性以刻画算法公平性，并利用假设检验方法验证算法公平性，最后利用零知识证明技术构建非接触式的信任。研究中利用滴滴出行数据就接驾距离验证了算法公平性，该实验验证了本方法的可行性及有效性，并展示了此框架用于更大规模、更完整的算法公平性检验的潜力。

1 引言

现今社会中存在大量算法，这些算法由平台企业开发，是平台向公众提供各种服务的基础，因此这些算法会对人们生活中如收入、福利等各个方面造成直接的影响。从而，这些算法的公平、透明、合理性都日渐受到监管者及公众的重视。特别的，网约车订单分配是平台算法的一个重要场景，订单分配算法直接影响了司机的收入和乘客的福利。因此订单分配算法公平性的重要性不言而喻，也是众多研究关注的方向。但是如何使监管者及公众相信算法的公平性仍是一个挑战。

在特定场景下，算法可以通过开源等手段实现透明、公开，从而构建信任。然而，在更多场景下，算法可能涉及平台的商业机密、数据涉及用户隐私，均不能直接公开。在此类场景下，密码技术是构建可信算法的传统方法之一，但是将算法计算过程逐步转为密文计算虽然能够保证结果可信，但是将引申出巨大的时间及计算成本。这造成了平台算法的信任困境。

为此，本文提出了基于统计的算法公平性检验方法，引入基于密码学的可信监管机器人扮演公平第三方，帮助企业和监管者与公众建立信任。具体而言，我们从统计学的角度利用随机性建模公平性，并利用假设检验方法检验算法的公平性，大幅降低计算成本。同时引入可信监管机器人作为第三方，利用加密承诺及零知识证明技术实现监管者与企业之间非接触式的检验方法，让监管者可以在不用深入算法细节、不用见到真实数据的前提下完成可信的算法公平性检验。

本文后续组织如下：第二章梳理了相关文献和工作，第三章介绍非侵入式的算法公平性检验理论，第四章展示了非侵入式的算法公平性检验方法，第五章呈现了一个实际的案例，第六章为结论。

*These authors contributed equally to this work.
†Corresponding Author(邮箱：yangyu1@mail.tsinghua.edu.cn,duruihuan@didiglobal.com)
‡清华大学人工智能国际治理研究院
§综合交通运输大数据处理及应用技术交通运输行业研发中心（中路高科交通科技集团有限公司）
¶滴滴出行

2 文献

算法的公平性是算法可信中重要的一个维度。Toreini 等人 [11] 将社会科学中的信任概念建模为计算机学科中的公平性、可解释性、可审计性以及安全，并提出了完整的算法可信框架。此外，也有大量探讨算法可信的文献中都谈及了公平性 [1][10]。

现有的算法公平方面的研究通常假定度量公平性的计算是由对数据具有完全访问权限的一方在本地完成的 [3][4][14]，基于分布式协议的算法中也以公平性为目标之一 [16]。然而，这两类方法对公平性都缺乏可验证性，因此不能保证算法可信。在特定场景中，具有可解释性的算法可以通过透明的设计而实现算法可信 [15][5][7]，但是这类方法只适用于特定的模型和场景，因此无法推广。此外，算法和数据在现实中是公司的重要资产，因此黑盒审计 [5] 是进行公平性检验的另一种方法。但是目前仍缺乏成熟的黑盒检验技术。

"公平性验证服务" 的概念以前已经被提出，例如在 [13] 中。作者们提出了一个框架，其中包含一个可信第三方作为算法公平性的担保人，负责进行公平性验证的计算，在这个框架中，所有利益攸关方都必须信任这个担保人。特别地，算法开发者必须将算法结果、敏感输入数据甚至模型的内部参数发送给担保人。这要求算法开发者相信担保人不会滥用这些信息，这个条件使得这个方法的实际应用受限。

为了解决这些限制，学者们开始在这类系统中引入加密技术，例如 Kilbertdus 等人 [6] 提出了一个名为 "blind justice" 的系统，利用多方安全计算协议，使用户（数据所有者）、模型（算法所有者）和监管者（验证公平性者）合作，使用联邦学习方法 [15] 进行公平的训练。由于这个方法对模型训练的过程进行加密，因此它依赖于模型和算法。Segal 等人 [9] 同样利用了加密技术进行公平性的计算于验证。Toreini 等人 [12] 则提出了利用零知识证明的公平性验证服务框架。而 Park 等人 [8] 为公平性计算提出了一个可信执行环境（TEE）。利用密码学上安全的特殊硬件组件，确保代码的正确执行。

3 非侵入式的算法公平性检验理论

网约车平台中，订单分配算法作为其提供服务的基础设施，撮合司机与乘客，其中，司机服务算法指派的订单从而获得收益，而乘客享受服务的质量如等待时间等也取决于订单分配算法所指派的司机。过程中，订单分配算法会影响司机与乘客的权益，因此算法的公平性至关重要。在传统的场景中，对于复杂度低的规则我们可以通过侵入式，即评估及验证规则本身以验证公平性，但是对于基于算法建立的复杂规则，如网约车订单分配算法，侵入式的检验复杂度极高，将带来高昂的时间及计算成本，且侵入式的检验需要利用的算法和数据可能涉及商业机密和用户隐私。为此，本研究提出一个非侵入式的网约车订单分配算法的公平性检验方法，无需深入算法细节，只需利用订单分配的场景信息和订单分配结果便能进行算法公平性检验，大幅降低计算成本，并能保护企业机密及用户隐私。

然而，非侵入式的算法不直接接触订单分配算法与数据，因此需要相应技术以保障用以检验的数据、计算过程与结果没有造假。为此，非侵入式公平性检验方法中引入了可信监管机器人作为公正第三方，利用非对称加密技术实现数据的可算不可见、透明但不公开，帮助监管者与平台企业构建非接触式信任。

本节包含算法公平性检验的理论以及利用非对称加密工具建立非接触式信任的监管机器人的理论。

3.1 订单分配算法的公平性检验

在网约车订单分配算法中，订单分配是在给定时间的乘客订单和司机信息下将订单指派给司机，其形式化地定义如下：对 N 个乘客订单 $\{P_i = (O_i, D_i, T_i)\}_{i=1\ N}$ 和 N 个司机 $C_i, i = 1\ N$，一个订单分配是由乘客订单到司机之间的一个双射 f，对于一个乘客订单 P_i，若 $f(P_i) = C_j$ 则表示司机 C_j 负责接送订单 P_i 的乘客。其中，O_i, D_i 分别代表乘客订单 P_i 的起始点和目的终点，而 T_i 则是乘客订单 P_i 的出发时间。

为了检验算法的公平性，侵入式的检验是直接评估算法的优化目标并验证其计算过程，然而多数情况下算法是平台的商业机密，且逐步检查算法的计算会带来巨大的成本。此外，逐步检查的做法依赖于具体算法，即检验不同的算法时，相应的计算过程也会随之变化，因此这样的检验方法不具有好的通用性。相比而言，非侵入式检验方法基于一个基本思想：公平的算法不应该系统性的置特定司机或用户于不利。由于算法给出的每次订单分配结果本身具有随机性-每次订单分配中，不同司机的接单距离及等待时间等指标必定存在差异，从而可能对司机的收益造成影响。因此，非侵入式检验方法通过研究算法衍生出的随机性以检验算法的公平性。

本节首先对订单分配算法的公平性进行刻画，具体而言，本节从统计学的视角利用算法的随机性进行建模，并论证在此之上进行公平性检验的可行性。其次，本节探讨了实际上统计推断的可检验性及可检验的条件。

3.1.1 调度算法的公平性理论

本节着重讨论订单分配算法的公平性与随机性的关系。在此之前，首先探讨订单分配算法影响司机收益的机制：订单分配算法在每次订单分配中会直接决定司机的接单距离、等待时长等因素，这些订单分配算法可控的因素和其他如司机出车时间、出车区域等订单分配算法不可控因素将共同决定司机的收入。本节假设所有算法不可控的因素一致，在此前提下讨论公平性和随机性的关系。

任何算法在每次给出的订单分配中，每个司机的接单距离及等待时间等指标都必然存在差异，因为在一次订单分配中无法保证每个司机的接单距离、等待时间等因素均完全一致，这便是算法的随机性。因此算法的公平性不应该按每次订单分配评估，而应该考虑一段时间内的长期效应，即一个公平的算法不应该影响司机长期的收入机会。具体而言，对一个订单分配 f，记司机 C_i 在这个订单分配下的收益为 $r_i(f)$。对司机 C_1 和 C_2 而言，即使一个公平的算法可能在某此订单分配 f 中使两个司机的收益 $r_1(f)$ 和 $r_2(f)$ 相差较大，但是长期而言，公平的算法在多次订单分配 $f_1, f_2, f_3, ...$ 中则应该使得这两个司机收益的期望接近，即 $|E(\sum_i r_1(f_i)) - E(\sum_i r_2(f_i))| < \epsilon$，其中 ϵ 是一个较小的数。

更严格地，上述性质的一个充分条件是体现在订单分配算法可控的因素上的随机性对不同司机而言是同分布的。这意味着每个司机在每次订单分配中接单距离、等待时间等影响收入的因素均服从相同的分布，实际的差异是由纯粹的随机性造成的，并非算法系统地置某个司机于不利，而算法的公平性是“机会公平”的。因此，非侵入式算法公平性检验便通过检验算法可控因素的随机性是同分布的以验证算法的公平性。

为了进一步体现订单分配算法的随机性，本节给出一个简单的线性模型上严格的理论推导：

考虑单一一个直线道路上有 N 个完全相同的司机与 N 个乘客，他们的位置是独立的均匀分布在道路上。下面通过理论建模并推导出最小距离订单分配算法下的随机性。其中，最小距离订单分配算法会选择使得所有司机去接乘客的总距离最短的一个订单分配。特别地，在单一直线道路上，对所有 $1 \leq k \leq N$，最小距离订单分配算法会指派离道路一端第 k 近的司机接送离道路同

一端第 k 近的乘客。

为了简单起见，在这个例子中只考虑一个指标-司机接单距离作为影响司机收益的因素，其中司机接单距离指的是司机由其初始位置到乘客出发位置的距离。

下面计算在此模型中司机接单距离的分布。

首先不妨假设直线道路的长为 1，则离道路一端第 k 近的司机的位置分布函数即均匀分布的第 k 个次序统计量的分布 $f_{(k)}(x) = \binom{N}{k-1} x^{k-1}(1-x)^{N-k}$ 服从 Beta 分布 $\beta(k, n-k+1)$。类似地，离道路一端第 k 近的乘客的位置分布函数 $g_{(k)}(x)$ 也服从相同的 Beta 分布 $\beta(k, n-k+1)$。

则离道路一端第 k 近的司机的接单距离满足分布

$$h_k(z) = 2\int_z^1 f_{(k)}(x)g_{(k)}(x-z)dx = 2N^2\binom{N-1}{k-1}^2\int_z^1 (x(x-z))^{k-1}((1-x)(1-x+z))^{N-k}dx \quad (1)$$

而对于一个司机而言，他离道路一端开始计算的排序是随机的，且位于各个顺序的概率均相等，因此他的接单距离分布为和

$$h(z) = \frac{1}{N}\sum_{k=1}^{N} h_{(k)}(z) \quad (2)$$

可见即使是没有恶意的算法，在理论推导下，也势必造成小部分的司机收益与其他司机不均等，这是由于算法的随机性造成的。

而公平的算法应该使所有司机的收益机会在长期是均等的。而通过验证算法影响司机收入的因素均是来自同样的分布便保证了其收益的机会的公平性。

3.1.2 可检验性

前节的讨论中假设了所有算法不可控的因素一致的前提。而在一般情况中，进行公平性检验之前需要对司机进行分类，以保证同类别中的司机算法不可控的因素一致，再检验算法对各类司机的公平性。

然而，并非对任何指标与分类方法总有办法进行验证。直观而言，若分类之后同类司机过少，将使得样本数不足，从而假设检验的置信度低；反之，为了保证一个类别中的司机数量够多可能使得分类条件过于宽松，从而导致其分布天然不应该满足公平性条件。由此可见，如何判断分类是否合适以及判断是否存在合适的分类是一个重要的问题，因此本节将讨论分类的颗粒度及指标的可验证性。其中，颗粒度指的是分类的标准，颗粒度大是分类宽松，颗粒度小是分类严格。

为了严格刻画上述现象，以下提出颗粒度的条件上限和颗粒度的样本下限两个概念：

- 颗粒度的条件 $\delta-$ 上限 U_δ: 给定的 δ 下，满足对任何两个同类的司机，算法不可控的因素差距不超过 δ 的最大颗粒度。

- 颗粒度的样本 $n-$ 下限 L_n: 给定的 n 下，使得同类司机的样本数不小于 n 的最小颗粒度。

对于一个指标而言，只有当 $U_\delta \geq L_n$ 时，其公平性才具有可检验性。如图 1、2 所示，对于具体的假设检验，不同深浅的蓝色样本固定颗粒度-p 值曲线表示在不同样本数量（n）的情况下，分类颗粒度越大会导致算法不可控因素的差距越大，从而可能导致假设检验中的 p 值较大，从而可能使结果不具统计意义。而绿色的实际颗粒度-p 值曲线线对应的是实际情况中，分类颗粒度小可能导致样本量小，造成假设检验的置信度低，即对应 p 值较大。因此，对于给定的 p 值，如 0.05，直线 $p = 0.05$ 和实际颗粒度-p 值曲线的交点对应的颗粒度 L 即是这个置信度要求下的样本下限 L_n，若颗粒度小于 $L = L_n$ 将会导致样本过少而使结果统计意义不满足要求。同时，直线

$p = 0.05$ 和样本固定颗粒度-p 值曲线的交点对应的颗粒度 U 便是这个置信度要求下的条件上限 U_δ，若颗粒度大于 $U = U_\delta$ 将导致类别中司机的算法不可控条件差距过大，无法通过检验。

图 1 中，$U = U_\delta = U > L = L_n$，因此对应的情况可检验，具体而言，我们可以选取满足 $L \leq F \leq U$ 的 F 作为检验中的分类颗粒度。而图 2 中，$U = U_\delta = U < L = L_n$，因此对应的情况不可检验，即不存在合适的分类颗粒度使得检验结果具有统计意义。

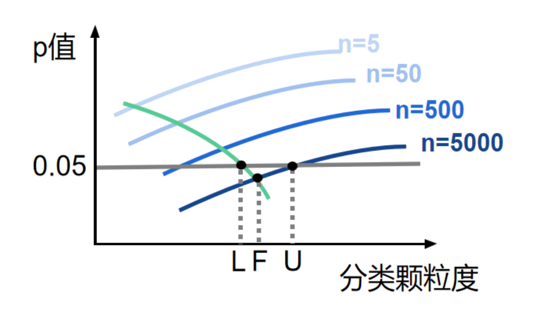


图 1: 可检验示意

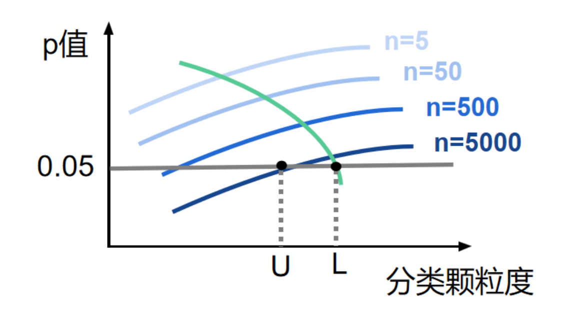


图 2: 不可检验示意

3.2 非接触式信任

非接触式信任的目标是使验证方在不用接触平台企业的算法和数据的情况下信任检验结果。为构建非接触式信任，我们在前述公平性检验的基础上引入利用加密承诺及零知识证明的技术构建的可信监管机器人，在保护数据安全的条件下确保检验结果正确。

直观的说，可信监管机器人由平台企业和验证方共同设置，此设置过程将保证双方无法进行潜在的作弊或攻击行为，从而建立信任。设置完成后，平台企业在需要验证时将数据发送给可信监管机器人，可信监管机器人基于先前的设置和系统记录的日志数据（包含企业机密及用户隐私）计算验证结果并生成证明，并将结果和证明发送给验证者，验证者只要验证证明便能够确认结果的正确性。在上述过程中，平台企业和验证者没有直接的交互，在可信监管机器人作为公正第三方的帮助下建立的非接触式信任，完成公平性验证。

具体而言，前述公平性检验使我们不用深入算法细节便能验证算法公平性，保护了平台的订单分配算法机密。而我们利用加密承诺以隐藏具体的数据，同时保证了平台无法篡改数据，因为加密承诺中，平台计算机密数据的 hash 值并将其公开作为承诺，若平台篡改了原始数据，将导致其无法通过承诺的验证，从而验证者在没有见到数据的情况下可以相信数据的真实性。最后，我们利用零知识证明保证验证计算结果的正确性。零知识证明使得验证者可以在无法获取或推导出企业机密及用户隐私的情况下，通过验证证明者发送的证明以确认公开的计算结果的正确性。

因此，通过基于统计的公平性检验方法、加密承诺和零知识证明的可信监管机器人，我们便能够构建平台与验证者直接的非接触式信任。

4 非侵入式的算法公平性检验方法

4.1 指标选取与司机分类

前两节中将验证订单分配算法公平性转化成计算关于指标的假设检验，本节讨论如何决定具体需要检验的指标，为此我们建立一个指标选取的方法体系，这个体系以确定检验环节和决定具

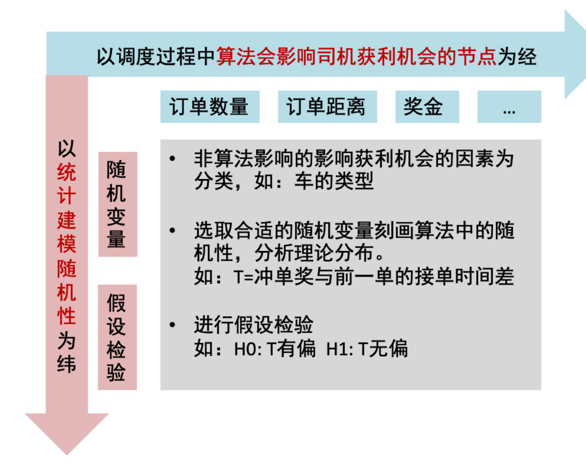


图 3: 指标方法论

体指标两步为经纬（图3），具体来说，首先是通过梳理网约车平台会影响个人盈利的节点确定哪些环节需要构建指标，然后则是在具体的环节通过统计建模刻画随机性，并设计合适的假设检验与指标。

4.1.1 建立个人盈利模型

本节介绍指标体系的“经”，通过梳理所有会影响个人盈利的节点或因素来建立个人盈利模型，并从中选择网约车平台会影响的部分确定哪些环节需要构建指标。

个人盈利模型中的影响因素分为两类，一类是算法能影响到获利机会的因素，另一类是算法不能影响，但是需要被考虑的个体本身分类。

例如，网约车场景下，司机服务的订单里程和服务时段都会影响司机收入，订单里程直接影响司机收入，不同时段乘客发出订单数量不同，也会间接影响司机收入，例如通勤高峰期订单量一般高于其他时段。每个司机服务的订单里程是网约车平台订单分配算法有机会影响的部分，而司机选择在什么时段接单是司机个人意愿，网约车平台算法不能影响。

4.1.2 设置分类并选取指标

确定了检验环节后，本节介绍指标体系的“纬”，即决定具体指标。

我们把影响因素中，算法无法影响的部分作为分类维度，可以影响的部分作为指标。在不同分类维度组合的类别内，对于指标选取合适的随机变量刻画算法中的随机性，并分析理论分布，然后进行假设检验。

例如网约车场景中，司机的接单等待时间（指标）会影响司机收入，而它既可以受算法影响，也可以受司机的服务时段等司机个体选择（分类维度）影响，所以在选择指标的时候，我们需要约束分类维度，在分类维度相同的情况下设计指标。

4.2 基于随机性的公平性订单分配假设检验构造

在前述例子中我们揭示了订单分配的结果由于随机性会造成司机收益机会的差异。因此，为了证明算法的公平性，我们构造假设检验以检验司机在一段时间范围内收益机会的差异是源于统

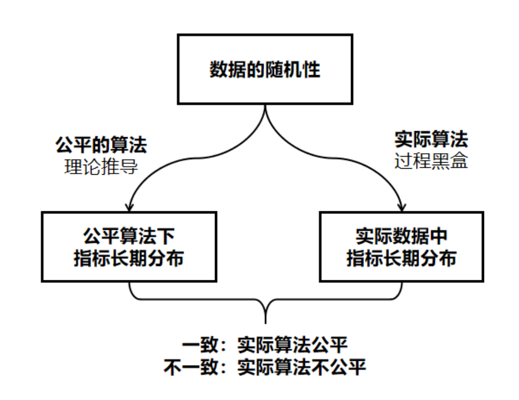


图 4: 算法公平性假设检验原理

计随机性，即算法在这一段时间范围内所有司机的收益机会都是均等的。我们延续前一节中的讨论，以对某个指标的验证为例。

如上一节的讨论，对于一个指标而言，由于算法随机性，势必造成不同司机在一轮订单分配中的指标取值有所差异。因此，我们需要利用统计来刻画这种随机性。假设在公平的订单分配算法下，同类司机的指标取值在每一轮中应该都是采样于同样的分布。我们将利用这个分布来检验算法的公平性。具体而言，对于某个指标，我们首先求出这个指标在一轮订单分配中的分布 $\mathcal{D}_s$，其次再计算多轮后该指标均值的理论分布 $\mathcal{D}_t$，并通过假设检验将实际分布 $\mathcal{D}_a$ 与理论分布 $\mathcal{D}_t$ 比较，从而验证算法公平性。本节中分为两步介绍构造公平性订单分配假设检验的方法。

4.2.1 指标在一轮订单分配中的分布

为了得到指标在一轮订单分配中服从的分布 $\mathcal{D}_s$，我们有基于理论和基于数值的两种方法：

- 基于理论的方法：如果假设订单分配算法足够简单，或可以将实际的订单分配算法简化为足够简单的算法，使得我们可以根据司机与乘客状态的分布理论的推导出在订单分配算法下指标的分布。例如上面的例子中，最小距离订单分配算法在司机与乘客的状态满足特定分布的情况下就可以由理论推导指标的理论分布。

- 基于数值的方法：在订单分配算法或司机与乘客的状态分布较复杂的情况下，我们可以通过数值方法确认指标在一轮订单分配中的分布。具体而言，我们可以选取历史上一个足够长的时间窗口，将这个指标的在这段期间内的频次作为指标在一轮订单分配中的分布。直观而言，这个方法也可以视为对某个司机而言，每一轮都是在所有同类司机中“抽样”。

4.2.2 多轮的分布与假设检验

直观上，在确定了每一轮的分布之后，我们一方面由理论可以推断多轮后司机指标均值的分布 $\mathcal{D}_t$，另一方面，我们可以由实际数据得到指标实际的分布 $\mathcal{D}_a$。因此我们可以通过假设检验方法检验实际的分布是否与理论分布相符。在订单分配算法公平的情况下，指标的实际分布将会贴近理论分布，从而通过假设检验；反之，在订单分配算法不公平的情况下，将没有理论保证实际分布贴近理论分布，除非订单分配算法被故意地“精心设计”，否则不公平的订单分配算法将无法通过假设检验。

为了使得检验方法具有良好的通用性，我们利用中心极限定理：独立同分布及方差有限的随机变量的的系列均值会服从正态分布。我们根据原始分布的期望和方差理论推导其最终收敛至的正态分布，并针对多轮的结果检验其正态性是否与理论相符。由此，我们对于不同分布的检验都可以被约化为正态性检验，从而大幅增加了此方法的通用性。

4.3 可信监管机器人系统设计

基于上述理论依据，本章介绍用以分析、检验和证明网约车订单分配算法公平性的可信监管机器人系统的设计。首先介绍系统中的各个参与者的定位，然后说明系统的目标，接着详细介绍系统的运行流程，最后分析系统是如何保证目标实现的。

4.3.1 系统参与者介绍

系统的参与者包含：证明者、验证者和机器人。三方的关系是，系统中有一个证明者，和一个（或一些）验证者，在双方不具备信任基础的前提下，证明者希望向验证者证明某个声明正确；同样的验证者也希望验证证明者的声明正确性；可信系统作为一个第三方（数字机器人）的角色帮助双方构建信任。

- 证明者：由企业担任，证明者公开声明事项，例如企业作为证明者证明订单分配算法的公平性，并和机器人交互提供证明所需输入。
- 验证者：验证者可以是主管部门、大众或用户等，验证者共同公开监管事项，例如司机作为验证者验证企业的订单分配算法公平性，通过和机器人交互得到证明者的数据是否支持声明正确的结论。
- 机器人：利用非对称加密技术构建的可信的监管机器人，其内部使用了多方安全计算、零知识证明以及加密承诺技术，其角色是透明的第三方，作为信任提供者与证明者验证者交互。

4.3.2 系统目标

为了支持算法公平性验证，监管机器人的系统需要满足以下几个目标：

- 正确性：正确性意味结果可信，即验证者在与系统交互后，信任系统得出的通过与否的结果。
- 零知识性：零知识性意味过程中系统记录的日志数据（包含企业机密及用户隐私）得到保护，即验证者在与系统交互后，无法得出或推断证明者的数据。

具体而言，我们利用非对称加密技术以实现以上目标，接下来首先介绍监管机器人系统的设计与流程，再说明这个系统如何保证正确性及零知识性。

4.3.3 系统流程

系统证明流程如图5所示，其中数据承诺为加密承诺技术，机器人内部为零知识证明技术，本系统使用 Groth16 算法，并利用 Circom[2] 编写电路，下面对各个步骤分别介绍：

1. 电路编写

证明者公开声明、验证者公开监管事项之后，二者公开假设检验方法，为了证明声明正确性，即证明可以通过假设检验，需要把假设检验的代数计算表示成等价的数字电路并公开，数字电路作为机器人的代数电路输入。图6展示了电路的两数相乘模块例子。

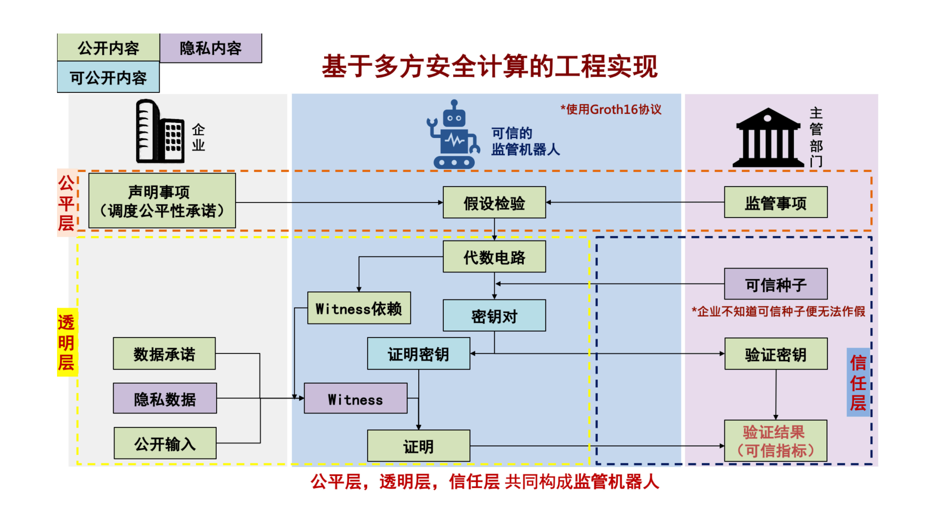


图 5: 系统流程

```
template Multiplier2(){
   //Declaration of signals
   signal input in1;
   signal input in2;
   signal output out <== in1 * in2;
}
```

图 6: 电路编写举例

编译电路可得到证明所需程序，分别是密钥对和 Witness 依赖，二者均可公开。密钥对在验证证明时使用，Witness 依赖证明生成时使用。

2. 初始设置验证者可以通过设置可信种子以保证企业证明过程中没有对数据和计算过程造假，因为企业不知道可信种子便无法作假，也因此可信种子需要对验证者以外的人保密。

根据可信种子和密钥对可以得到证明密钥和验证密钥，二者均可公开。证明密钥在生成证明时使用，验证密钥在验证证明时使用。

3. 生成证明（Groth16 协议）

证明者向机器人输入系统记录的日志数据（包含企业机密及用户隐私）作为机器人的数据输入，该数据只有证明者和机器人可见。

在产生底层数据的时候还要对对数据进行加密承诺（数字指纹），即计算数据的承诺值，并提交给机器人，验证阶段使用数据时，计算并验证数据承诺，如果一致即可证明数据没有受到篡改。同时数据承诺为单向的，由数据可以计算数据承诺，但是反之不行，所以这一步也不会泄露隐私。该数据是公开的。

另外还可以输入一些可以公开的数据，作为机器人的公开输入。

根据证明者的数据输入和编译电路得到的 witness 依赖，得到 witness，witness 是零知识证明中证明者数据的见证者，一组数据对应一个 witness，从 witness 中可以知道证明者数据，是零知识证明需要的输入形式，该数据验证者不可见。

根据证明密钥和 witness，可以生成证明，这一步可以保证企业的数据不公开，但透明，即这个证明已经可以公开了，不会泄露证明者系统记录的日志数据（包含企业机密及用户隐私）。

4. 验证证明（Groth16 协议）

使用验证密钥去验证证明，即可得到证明结果。如果通过证明，就表示证明者提供的数据通



图 7: 司机和乘客视角下派单流程

过了声明的假设检验。

4.3.4 系统分析

本节分析上述监管机器人系统如何保证正确性及零知识性。

- 正确性：结果的正确性由数据的真实性和计算过程的正确性保证。数据的真实性意味着证明者输入的数据真实，从而证明者无法通过输入不真实的数据造假。这由加密承诺保证，
- 零知识性：零知识证明保证了通过验证证明，验证者只能得知关于计算结果的信息，而无法获取或推导出企业机密及用户隐私，这便保障了此系统的零知识性。

5 实现案例

基于理论依据和系统设计方法，本章以一个网约车场景的具体实现案例为例，首先介绍司机个人盈利模型的建模思路和模型，然后介绍一个具体指标的选取、分布，以及假设检验方法，最后介绍证明系统在该指标上的结果与表现。

5.1 一个简单的个人盈利模型

在网约车场景下，分别从司机和乘客的视角下，简化版派单流程如图7所示，为了方便理解，图中以滴滴出行为例展示了司机端和乘客端界面。

司机打开司机端 app 后，点击"出车"表示开始等待接单，在接到订单之前，司机会按照喜好游走或者停车等待。网约车平台算法为司机找到订单（即司机接到订单）后，司机首先去接驾，即从目前位置去乘客上车点接乘客。接到乘客后，司机开始送驾，即从乘客上车点送乘客去乘客目的地，同时开始计费。送达乘客到目的地后司机完成订单，同时结束计费。完成订单后司机继续选择等待下一单，或者收车，即不再等待接单。

订单数 = 服务时长 /(等待时间 + 接送时间)	订单平均里程	平均每公里司机收入
• 接单时段 • 接单区域 • 服务时长 • 接单等待时长 • 订单接驾距离 • 订单送驾距离	• 订单平均里程	• 订单平均里程 • 车型

图 8: 收入建模

乘客侧流程类似，乘客打开乘客端 app 后，输入上车点和目的地等信息后，点击“确认呼叫”即发出订单。网约车平台算法为订单找到司机（即订单被应答）后，乘客等待司机接驾。乘客被接到后，司机送其去目的地。到达目的地后订单完成。

所以司机的收入可以建模为如下公式：

$$\begin{aligned}\text{司机收入} &= \text{订单数} * \text{订单平均里程} * \text{平均每公里司机收入} \\ &= f(\text{接单时段}, \text{接单区域}, \text{服务时长}, \text{接单等单时长}, \text{接驾距离}, \text{送驾距离}, ...) \\ &* \text{订单平均里程} \\ &* g(\text{订单平均里程}, \text{车型}, ...)\end{aligned} \tag{3}$$

其中，服务时长为司机出车服务的总时长。

5.2 选取指标的分布，假设检验结果

司机收入模型中（图8），影响司机收入的因素可以分成两类，一类（红色）是派单系统无法影响的，比如接单时段、接单区域、服务时长和车型；一类（绿色）是派单系统可以影响的，比如等单时长、接驾时长、送驾时长和订单平均里程。

我们把派单系统无法影响的部分作为分类维度，可以影响的部分作为指标。比如，根据接单时段、接单区域、服务时长和车型等对司机进行分类，对同分类司机的人均接驾距离进行假设检验。

假设订单维度（即每个样本是一个订单）接驾距离 pd_o 服从 g 分布，$pd_o \sim g$，我们定义单个司机的平均接驾距离 apd_c 就是每个司机 c 服务的 x_c 个订单的接驾距离的平均值。

$$apd_c = \frac{\sum_{i=1}^{x_c} pd_i}{x_c} \tag{4}$$

如果派单系统不干预同分类司机的接驾距离，则 pd_o 的分布与司机个体独立，在所有司机中的分布均为 g。那么，每个司机 c 的 apd_c 就是一次对 pd_o 的采样的均值，根据中心极限定理，如果 pd_o 独立同分布，且具有期望 $\mu = E(pd_o)$、方差 $\sigma^2 = D(pd_o) > 0$,x_c 足够大的时候，apd_c 服从正态分布。

$$apd_c \sim N(\mu, \frac{\sigma^2}{x_c}) \tag{5}$$

所以证明思路为：如果派单系统不干预同分类司机内部的接驾距离，那么分类因素确定时，接单数 x_c 足够大（大于 A）的 n 个司机 $cset = \{c \mid x_c > A\}$ 的平均接驾距离服从正态分布。需要

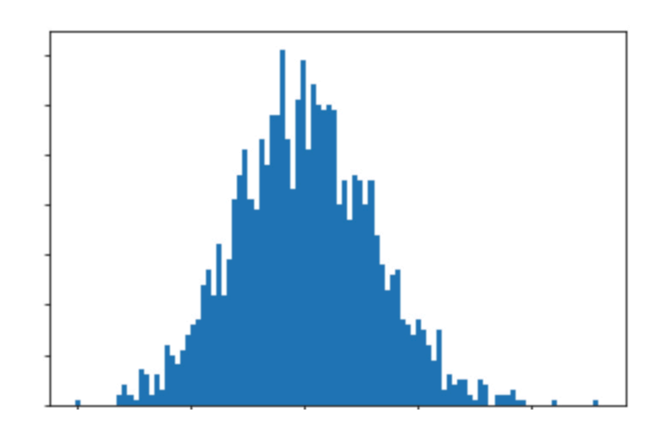


图 9: 某分类因素下，标准化后司机平均接驾距离分布

注意，这里需要对平均接驾距离标准化，因为 apd_c 的方差与 x_c 相关，不同 x_c 的司机的平均接驾距离服从不同的分布。标准化计算公式如下。

$$napd_c = \frac{apd_c - \mu}{\sigma/\sqrt{x_c}} \tag{6}$$

某些分类因素约束下，标准化后司机平均接驾距离分布如图9所示。

证明司机平均接驾距离服从正态分布，本文采用偏度-峰度检验法，正态分布的偏度 b_s 和峰度 k 都等于 0。

偏度检验时，假设为 $H_0 : b_s = 0, H_1 : b_s \neq 0$, 偏度检验统计量 $\hat{b_s} = \frac{m_3}{m_2^{3/2}}$, 峰度检验统计量 $\hat{k} = \frac{m_4}{m_2^2} - 3$, 其中 $m_i = \frac{1}{n}\sum_{c \in cset}(napd_c - \overline{napd_c})^i$，在 α 的置信度下，以下条件满足时不能拒绝无偏。

$$\begin{aligned} |\hat{b_s}| &\leq \sqrt{\frac{6(n-2)}{(n+1)(n+3)}} z_{1-\alpha/2} \\ |\hat{k} + \frac{6}{n+1}| &\leq \sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}} z_{1-\alpha/2} \end{aligned} \tag{7}$$

其中 $z_{1-\alpha/2}$ 查标准正态分布表可得。

本例中，算法不可控的因素差距为 5%, 同类司机的样本为 1.3k，满足可检验性。

5.3 证明系统结果与表现

本例中，把证明拆分成四个电路，如图10所示。其中，预处理电路验证从原始数据到电路输入数据的预处理过程，包括从订单维度信息聚合成司机维度信息、标准化、电路必须的放缩取整。偏度检验和峰度检验分别对电路输入数据进行偏度检验和峰度检验。加密承诺电路验证原始数据的加密承诺过程。

以偏度检验电路为例，证明系统流程图如图11所示，其中 bs_test.circom 文件为证明电路，该实例中大小为 4KB，bs_test.r1cs 为编译电路产物之一，1.4MB，proof.json 为证明文件，4KB，输入数据量为 1k+ 的时候，运行时间为 6 分钟。

电路1：预处理	电路2：偏度检验	电路3：峰度检验	电路4：加密承诺
输入： $pd_o, apd_c, x_c, \sqrt{x_c}, \mu, \sigma, napd_c, A, \overline{napd_c}, z$ 验证： apd_c计算 μ, σ计算 $\sqrt{x_c}^2 + 1 > x_c \& \sqrt{x_c}^2 - 1 < x_c$ $\bar{y}$计算 z计算	输入： z 验证： 偏度检验通过	输入： z 验证： 峰度检验通过	输入： apd_c, h 验证： $h = hash(apd_c)$

其中，x_c大于阈值A的$napd_c \xrightarrow{\text{放缩取整}} y \longrightarrow z = y - \bar{y}$ ，h是数据产生时计算的承诺值

图 10: 证明电路划分

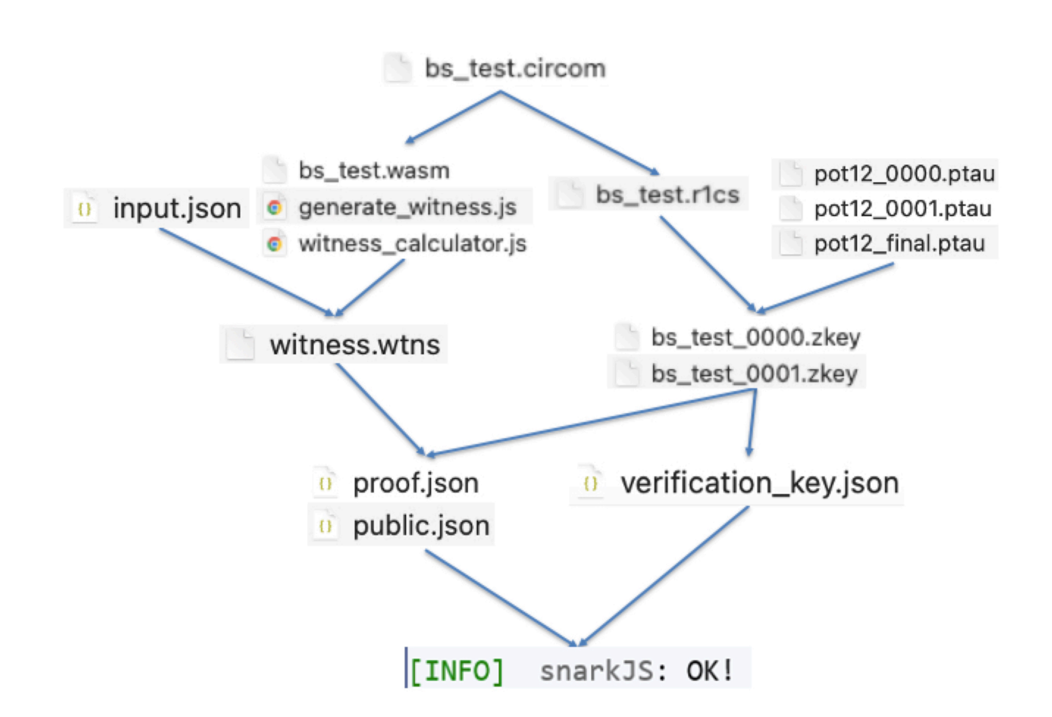


图 11: 证明系统流程图

6 结论

我们提出的基于统计的算法公平性检验方法可以有效解决算法的信任构建问题，通过我们提出的验证方法检验算法的公平性，并通过多方参与的安全计算验证确保结果可信。

本方法具有非接触式的特性，验证者无需深入算法细节、不用知道系统记录的日志数据（包含企业机密及用户隐私），便能完成算法公平性的验证。在非接触式的前提下，利用零知识证明技术确保结果可信，完成非接触式信任的构建。此外，通过统计建模公平性而非逐步验证算法计算细节，我们大幅降低了验证所需的计算与时间成本，在实验中，我们也能高效的完成公平性的检验、证明与验证。

本文提出的方法也是一个重要的模式创新，通过基于密码学的机器人作为可信第三方，帮助企业与主管部门和公众构建信任。这样的模式未来可以扩展至更多类似的场景中，帮助更多算法实现可验不可见、透明但不公开的非接触式信任。

值得一提的是，目前的检验体系利用基于假设检验的方法大幅降低计算与验证的成本，因此通过此公平性检验仅是算法公平的必要条件，即我们保证公平的算法能够通过检验，而对于不公平的算法，我们认为在极端情况下，若是被精心设计，也可能通过本文中的公平性检验。后续研究中可以通过添加更多必要条件的验证以进一步加强识别不公平算法的准确率。

参考文献

[1] Bacciarelli, A. The toronto declaration: Protecting the right to equality and non-discrimination in machine learning systems.

[2] Bellés-Muñoz, M., Isabel, M., Muñoz-Tapia, J. L., Rubio, A., and Baylina, J. Circom: A circuit description language for building zero-knowledge applications. *IEEE Transactions on Dependable and Secure Computing* (2022).

[3] Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data 5*, 2 (2017), 153–163.

[4] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (2017), pp. 797–806.

[5] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. A survey of methods for explaining black box models. *ACM computing surveys (CSUR) 51*, 5 (2018), 1–42.

[6] Kilbertus, N., Gascón, A., Kusner, M., Veale, M., Gummadi, K., and Weller, A. Blind justice: Fairness with encrypted sensitive attributes. In *International Conference on Machine Learning* (2018), PMLR, pp. 2630–2639.

[7] Panigutti, C., Perotti, A., Panisson, A., Bajardi, P., and Pedreschi, D. Fairlens: Auditing black-box clinical decision support systems. *Information Processing & Management 58*, 5 (2021), 102657.

[8] Park, S., Kim, S., and Lim, Y.-s. Fairness audit of machine learning models with confidential computing. In *Proceedings of the ACM Web Conference 2022* (2022), pp. 3488–3499.

[9] Segal, S., Adi, Y., Pinkas, B., Baum, C., Ganesh, C., and Keshet, J. Fairness in the eyes of the data: Certifying machine-learning models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (2021), pp. 926–935.

[10] Siau, K., and Wang, W. Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal 31*, 2 (2018), 47–53.

[11] Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., and Van Moorsel, A. The relationship between trust in ai and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (2020), pp. 272–283.

[12] Toreini, E., Mehrnezhad, M., and van Moorsel, A. Fairness as a service (faas): verifiable and privacy-preserving fairness auditing of machine learning systems. *International Journal of Information Security* (2023), 1–17.

[13] Veale, M., and Binns, R. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society 4*, 2 (2017), 2053951717743530.

[14] Wang, G., Zhong, S., Wang, S., Miao, F., Dong, Z., and Zhang, D. Data-driven fairness-aware vehicle displacement for large-scale electric taxi fleets. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)* (2021), IEEE, pp. 1200–1211.

[15] Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., and MacNeille, P. A bayesian framework for learning rule sets for interpretable classification. *The Journal of Machine Learning Research 18*, 1 (2017), 2357–2393.

[16] Zhang, Y., Zhou, Y., Hu, Y., and Huang, H. A decentralized ride-hailing mode based on blockchain and attribute encryption. In *International Symposium on Cyberspace Safety and Security* (2022), Springer, pp. 301–313.